MLUX 2020

Being Actively Ethical: Dynamic UX for AI

Carol J. Smith
Sr. Research Scientist - Human-Machine Interaction, CMU’s SEI
Adjunct Instructor, CMU’s Human-Computer Interaction Institute

Twitter: @carologic @sei_etc

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Copyright Statement

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM20-0795

Responsible, Intentional Design

Just because you can, doesn't mean you should.

Early, purposeful work

What is the challenge being face? Is it AI-friendly?

For whom? What are their needs?

What kind of improvements are expected?

What might a machine do better or faster?

What is not going to be improved (out of scope)?

AI is a partner - augmenting our abilities

Speed

- Find patterns
- Calculations

Safety (robotics)*

- Dull
- Dirty
- Dangerous
- Dear

*Marr, B. "The 4 Ds Of Robotization: Dull, Dirty, Dangerous And Dear." Forbes. Oct 16, 2017. https://www.forbes.com/sites/bernardmarr/2017/10/16/the-4-ds-of-robotization-dull-dirty-dangerous-and-dear/#70749ed83e0d

Diverse teams

Gender, race, culture

Education (school, program, etc.)
Experiences
Thinking process,
Disability status,
and more…

Photo by Christina @ wocintechchat.com on Unsplash
https://unsplash.com/@wocintechchat?utm_source=unsplash&amp;utm_medium=referral&amp;utm_content=creditCopyText

Not lowering bar
____________ extending it

Diverse, talented and multi-disciplinary

Includes skill set and problem framing approach

UX Professionals (big umbrella)

Data Scientists, Machine learning experts

Programmers, System architects

Product managers, etc.

Representatively diverse leadership for retention

Inclusive – Individuals' differences are acknowledged and accepted

Great Minds Think Different

High value in diverse teams

Focus more on facts

Process facts more carefully

More innovative

“…become more aware of their own potential biases”

Harvard Business Review

Why Diverse Teams Are Smarter. Harvard Business Review.
https://hbr.org/2016/11/why-diverse-teams-are-smarter

Ethics for Technology

AI has great potential, develop with caution

Future AI’s *may* be trusted to substitute human cognition and abilities.

Humans must continue to be responsible for situations that involve a person’s:

- Life (the use of force)
- Quality of life
- Health
- Reputation

“AI will ensure appropriate human judgement and not replace it” - DIB

To be biased, is to be human

Bias are shortcuts, to avoid risk and simplify problems.

Not inherently bad,
may be misapplied

Implicit = invisible

Not necessarily in sync
with our conscious beliefs

Can be managed and changed

Talk about biases in non-threatening, productive ways

Biased due to…

Social class

Resource availability

Education

Race, gender, sexuality

Culture, theology, tradition

More…

All systems have some form of bias

Complete objectivity is misleading.

Bias can have purpose and can be helpful.

The goal is to reduce unintended and/or harmful bias.

Adopt Technology Ethics

- Harmonize cultural variations
- Balance to pace of change, industry pressure
- Explicit permission to consider and question breadth of implications

acm Association for Computing Machinery

AI NOW INSTITUTE

Microsoft

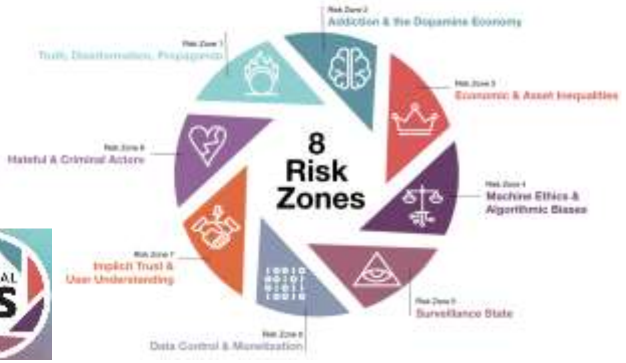
Google

Montréal Declaration Responsible AI_

An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

ETHICAL OS


8 Risk Zones
Risk Zone 1
Truth, Disinformation, Propaganda
Risk Zone 2
Addiction & the Dopamine Economy
Risk Zone 3
Economic & Asset Inequalities
Risk Zone 4
Machine Ethics & Algorithmic Biases
Risk Zone 5
Surveillance State
Risk Zone 6
Data Control & Monetization
Risk Zone 7
Implicit Trust & User Understanding
Risk Zone 8
Hateful & Criminal Actors

Coalesce on Shared Set of Technology Ethics

< >
Montréal Declaration
Responsible AI_
</ >

An initiative of Université de Montréal

1. Well-being
2. Respect for autonomy
3. Protection of privacy and intimacy
4. Solidarity
5. Democratic participation
6. Equity
7. Diversity inclusion
8. Prudence
9. Responsibility
10. Sustainable development

Montréal Declaration for a responsible development of artificial intelligence.
https://www.montrealdeclaration-responsibleai.com/the-declaration

Diverse, inclusive leaders

Diverse, Multi-Disciplinary Teams

Shared Tech Ethics

UX Framework

Designing Trustworthy AI

Activate curiosity

UX research methods to activate curiosity:

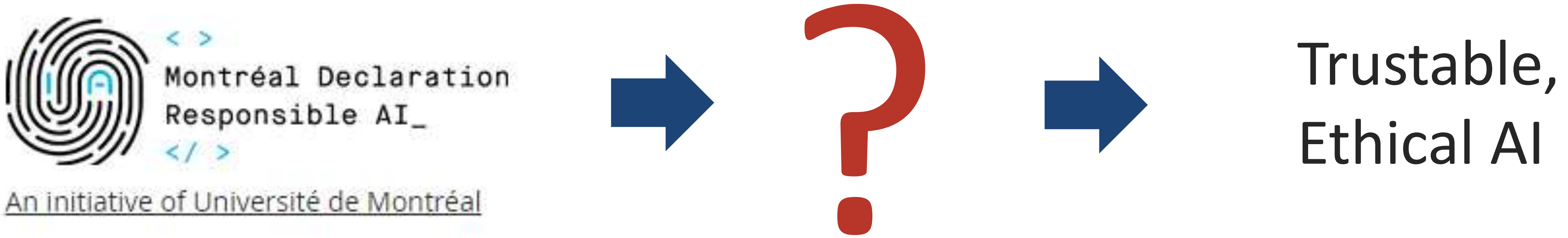
- Abusability Testing
- “Black Mirror” Episodes (inspired by British dystopian sci-fi tv series of same name)
- Flip it to test it
- Implicit Association Test from Harvard

Speculate about system misuse and abuse

- What are potential unintended/unwanted consequences?

More methods to “Outsmart Your Own Biases.”: https://hbr.org/2015/05/outsmart-your-own-biases
Implicit Association Test (IAT): https://implicit.harvard.edu/implicit/takeatest.html

How do we get there?



Montréal Declaration Responsible AI_

An initiative of Université de Montréal

Trustable, Ethical AI

Montréal Declaration for a responsible development of artificial intelligence. https://www.montrealdeclaration-responsibleai.com/the-declaration

Conversations for Understanding

UX Framework guides AI teams

Difficult Topics

- What do we value?
- Who could be hurt?
- What lines won’t our AI cross?
- How are we shifting power?*
- How will we track our progress?

*”Don’t ask if artificial intelligence is good or fair, ask how it shifts power.” Pratyusha Kalluri.
https://www.nature.com/articles/d41586-020-02003-2

Photo by Pam Sharpe https://unsplash.com/@msgrace?utm_source=unsplash&amp;utm_medium=referral&amp;utm_content=creditCopyText On Unsplash -
https://unsplash.com/s/photos/business-woman-smiling?utm_source=unsplash&amp;utm_medium=referral&amp;utm_content=creditCopyText

*“How is this ML model shifting power?" @riakall #NeurIPS2019

New uncomfortable work

"Be uncomfortable"

- Laura Kalbag

Ethical design is not superficial.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

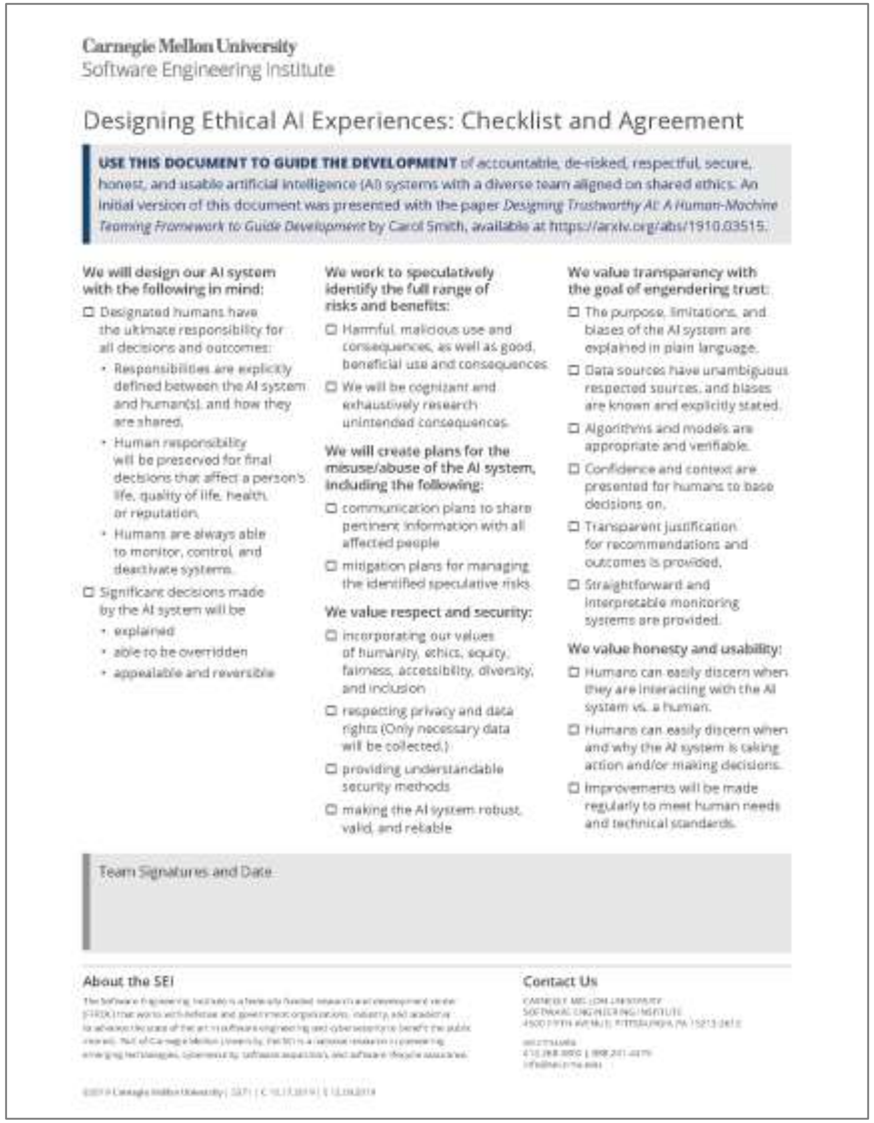
Prompt conversations

Pair Checklist with Technical Ethics

- Bridges gap between “do no harm” and reality

Reduce risk and unwanted bias

Support inspection and mitigation planning



Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, de-risked, respectful, secure, honest, and usable artificial intelligence (AI) systems with a diverse team aligned on shared ethics. An initial version of this document was presented with the paper *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development* by Carol Smith, available at https://arxiv.org/abs/1910.03515.

We will design our AI system with the following in mind:

□ Designated humans have the ultimate responsibility for all decisions and outcomes:
- Responsibilities are explicitly defined between the AI system and human(s), and how they are shared.
- Human responsibility will be preserved for final decisions that affect a person's life, quality of life, health, or reputation.
- Humans are always able to monitor, control, and deactivate systems.

□ Significant decisions made by the AI system will be
- explained
- able to be overridden
- appealable and reversible

We work to speculatively identify the full range of risks and benefits:

□ Harmful, malicious use and consequences, as well as good, beneficial use and consequences.

□ We will be cognizant and exhaustively research unintended consequences.

We will create plans for the misuse/abuse of the AI system, including the following:

□ communication plans to share pertinent information with all affected people

□ mitigation plans for managing the identified speculative risks

We value respect and security:

□ incorporating our values of humanity, ethics, equity, fairness, accessibility, diversity, and inclusion

□ respecting privacy and data rights (Only necessary data will be collected.)

□ providing understandable security methods

□ making the AI system robust, valid, and reliable

We value transparency with the goal of engendering trust:

□ The purpose, limitations, and biases of the AI system are explained in plain language.

□ Data sources have unambiguous respected sources, and biases are known and explicitly stated.

□ Algorithms and models are appropriate and verifiable.

□ Confidence and context are presented for humans to base decisions on.

□ Transparent justification for recommendations and outcomes is provided.

□ Straightforward and interpretable monitoring systems are provided.

We value honesty and usability:

□ Humans can easily discern when they are interacting with the AI system vs. a human.

□ Humans can easily discern when and why the AI system is taking action and/or making decisions.

□ Improvements will be made regularly to meet human needs and technical standards.

Team Signatures and Date

About the SEI

The Software Engineering Institute is a federally funded research and development center (FFRDC) that works with defense and government organizations, industry, and academia to advance the state of the art in software engineering and cybersecurity to benefit the public interest. Part of Carnegie Mellon University, the SEI is a national resource in pioneering emerging technologies, cybersecurity, software acquisition, and software lifecycle assurance.

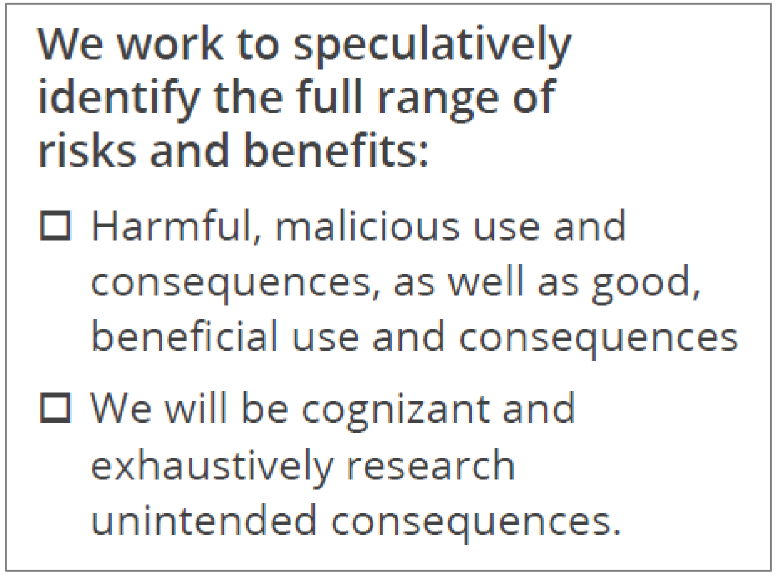
Contact Us

CARNEGIE MELLON UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE; PITTSBURGH, PA 15213-2612

sei.cmu.edu
412.268.5800 | 888.201.4479
info@sei.cmu.edu

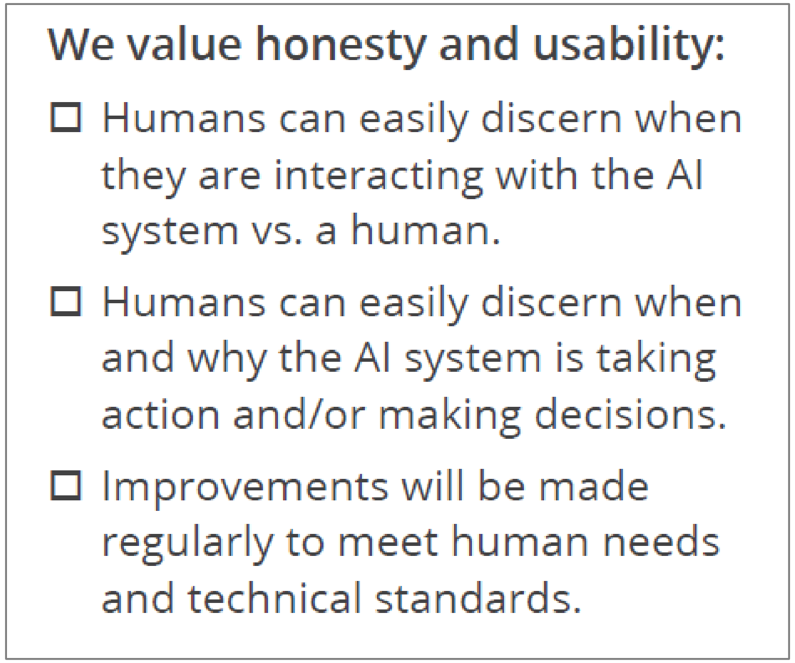
Checklist and Agreement - Downloadable PDF:
https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

Prompts help reveal hidden tasks



We work to speculatively identify the full range of risks and benefits:

- ☐ Harmful, malicious use and consequences, as well as good, beneficial use and consequences
- ☐ We will be cognizant and exhaustively research unintended consequences.

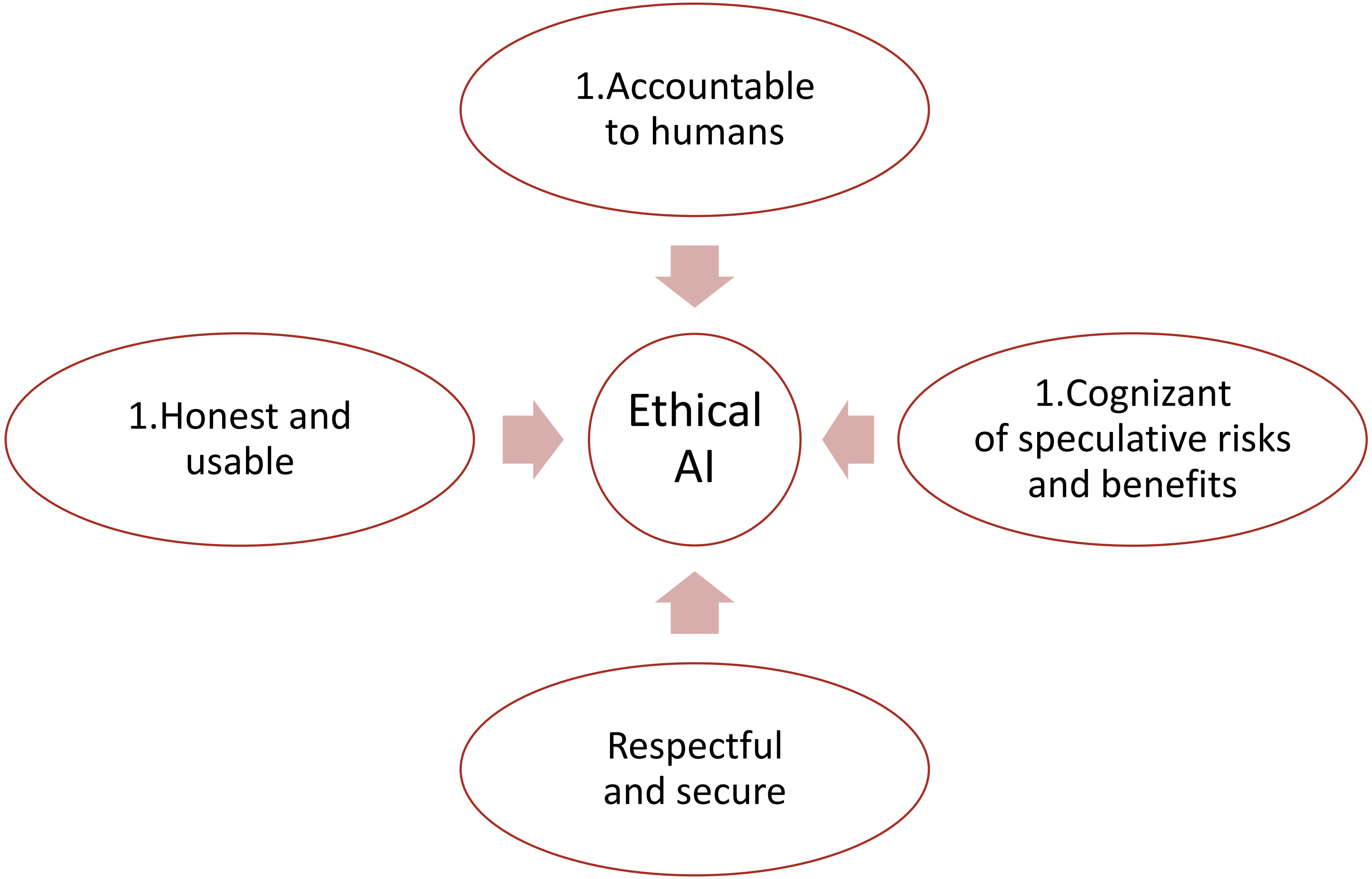


We value honesty and usability:

- ☐ Humans can easily discern when they are interacting with the AI system vs. a human.
- ☐ Humans can easily discern when and why the AI system is taking action and/or making decisions.
- ☐ Improvements will be made regularly to meet human needs and technical standards.

Checklist and Agreement - Downloadable PDF at SEI:
https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

UX Framework for Designing Trustworthy AI



RightStaff Scenario

AI shift scheduling system

Users: Store managers of fast food restaurants

Goals of RightStaff:

- Faster staffing decisions and scheduling
- Reduced bias of shift selection

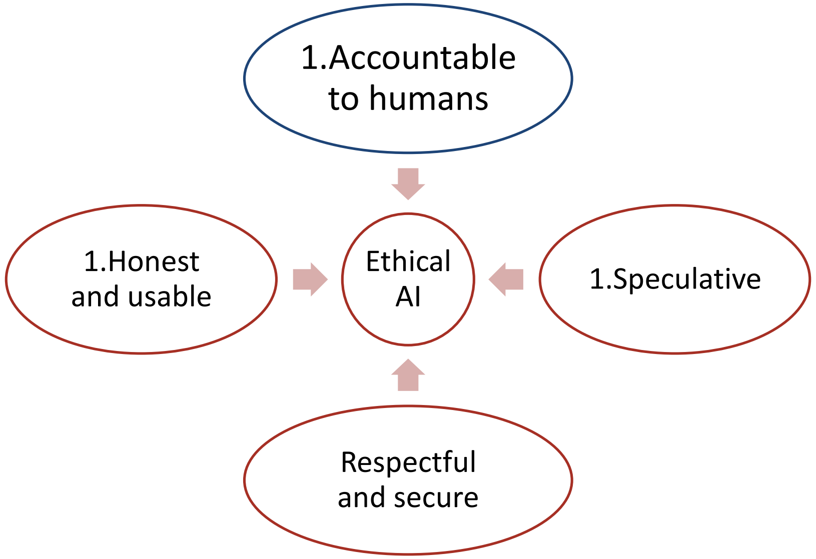
Accountable to Humans

Ensure humans have ultimate control

- Able to monitor and control risk

Human responsibility for final decisions

- Person’s life
- Quality of life
- Health
- Reputation



“Ensure humans can unplug the machines”

– Grady Booch

TED Talk, Grady Booch, Scientist, Philosopher, IBM’er
https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

Significant decisions

Significant decisions made by the AI system will be

- explained
- able to be overridden
- appealable and reversible

RightStaff

- Manager able to reschedule people as needed

Responsibilities explicitly defined

Between AI system and human(s)

RightStaff (AI System *or Manager?*)

- Picks employees to schedule?
- Defines shifts?
- Method to integrate new information?
  - Sick time
  - Resignations

Abusability Testing

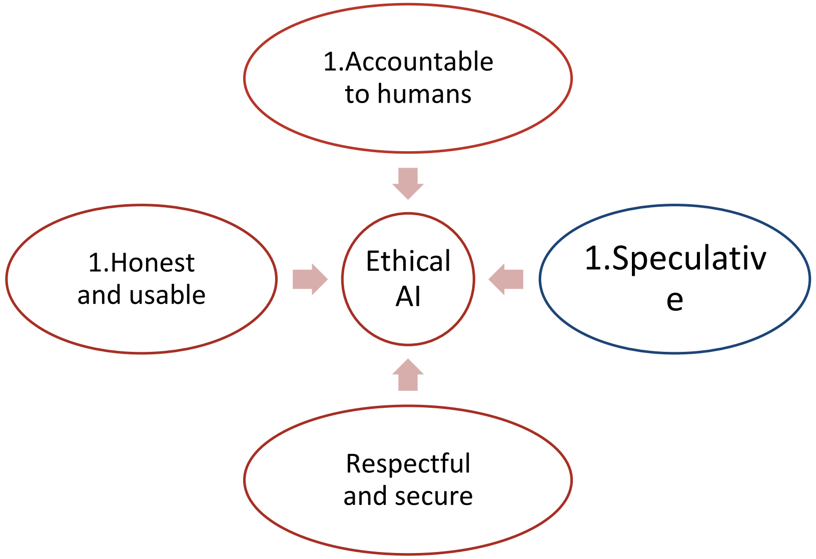
Feature added to enable RightStaff to turn off by itself

- What are limits to functionality?
- How could this be abused/misused?
- Implications?
- Risks?

Cognizant of Speculative Risks and Benefits

Identify full range of

- Harmful, malicious use, as well as good, beneficial use
- Blind spots and unwanted/unintended consequences



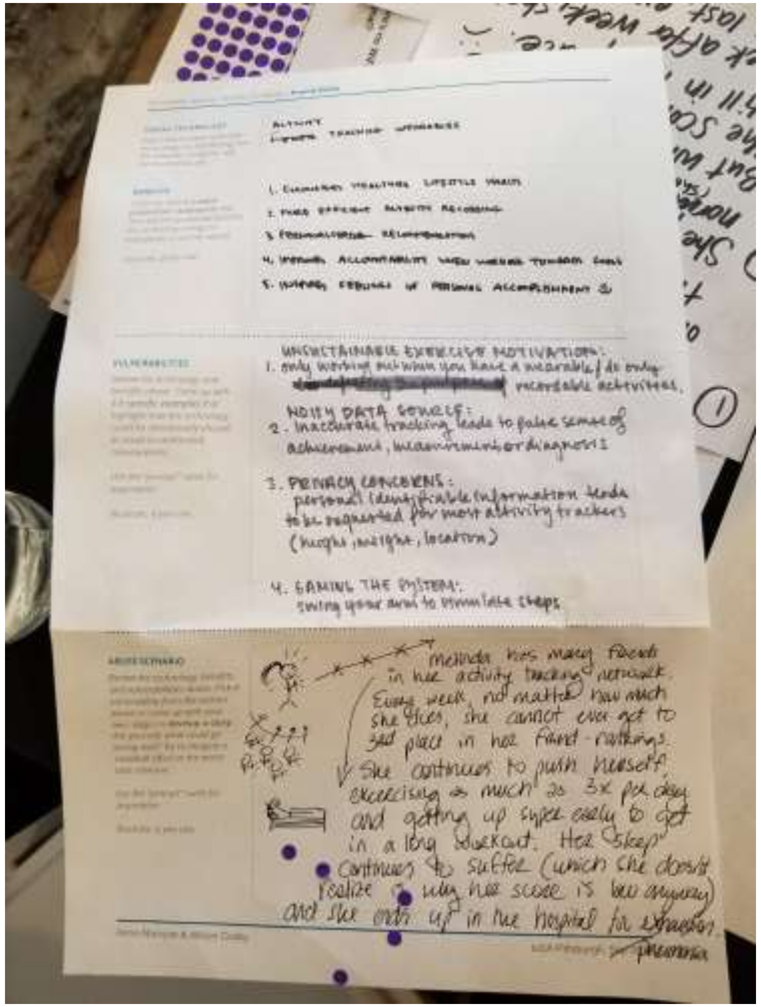
Speculative: Conduct UX research and activate curiosity

Speculate about misuse
and abuse

Potential severe abuse
and consequences

Perspective of people
in frequently marginalized groups

“Black Mirror” episodes



ACTIVITY
FITNESS TRACKING WEARABLES

1. ENCOURAGES HEALTHIER LIFESTYLE HABITS
2. MORE EFFICIENT ACTIVITY RECORDING
3. PERSONALIZED RECOMMENDATIONS
4. PERSONAL ACCOUNTABILITY WHEN WORKING TOWARDS GOALS
5. INSPIRES FEELINGS OF PERSONAL ACCOMPLISHMENT ☺

UNSUSTAINABLE EXERCISE MOTIVATION:
1. only working out when you have a wearable / do only [redacted] recordable activities.

NOISY DATA SOURCE:
2. Inaccurate tracking leads to false sense of achievement, misinformation or diagnosis

3. PRIVACY CONCERNS:
personal identifiable information tends to be requested for most activity trackers (height, weight, location)

4. GAMING THE SYSTEM:
swing your arm to simulate steps

Melinda has many friends in her activity tracking network. Every week, no matter how much she tries, she cannot ever get to 3rd place in her friend-rankings. She continues to push herself, exercising as much as 3x per day and getting up super early to get in a long workout. Her sleep continues to suffer (which she doesn't realize is why her score is low anyway) and she ends up in the hospital for exhaustion pneumonia

“Black Mirror” episode

- RightStaff begins prioritizing people with easier schedules
- Managers approve these schedules, reinforcing bias
- People who were previously discriminated against are *still* discriminated against

- What else?

Speculative: Create communication & mitigation plans

Plan for unwanted consequences

Misuse and abuse of AI system

- Who can report?
- To whom?
- Turn off?
- Who notified?
- Consequences?

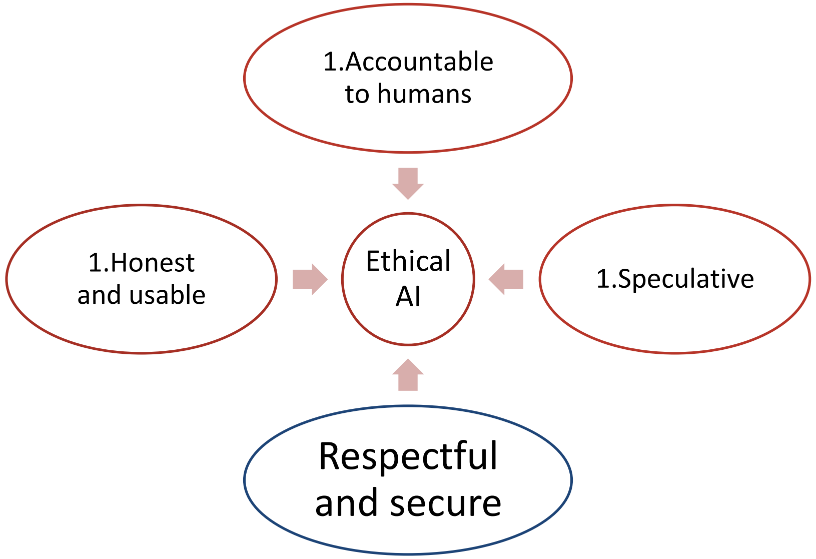
Respectful and Secure

Values of humanity, ethics, equity, fairness, accessibility, diversity and inclusion

Respect privacy and data rights

Make system robust, valid and reliable

Provide understandable security



Respectful and Secure

RightStaff

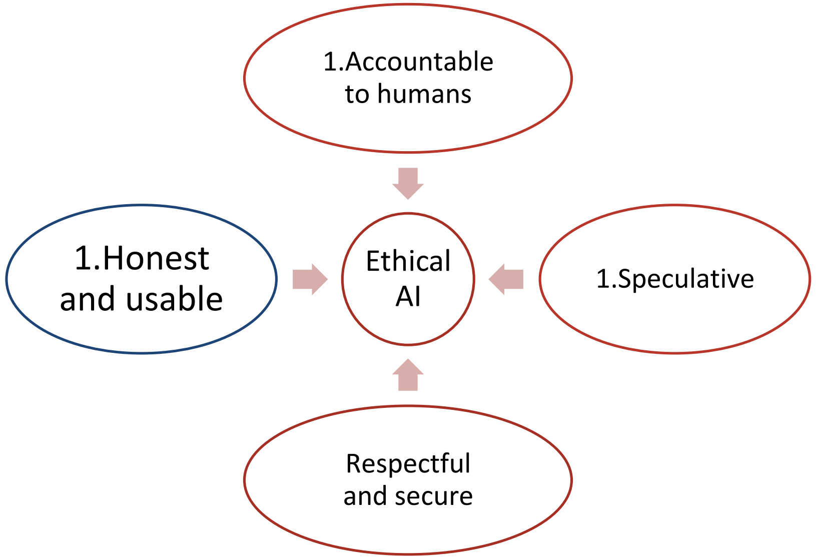
- Who has visibility to reasons for changing schedules?
- How is that information used?
- How is PII* of employees protected?

*PII is Personally Identifiable Information (social security number, address, etc.)

Honest and Usable

Value transparency with the goal of engendering trust

Explicitly state identity as an AI system



Fair: Remove unwanted bias in data

Show awareness of known and desirable bias

Acknowledge issues

Overcommunicate on issues

RightStaff

- System built to reduce the known bias in existing data
- Make it easy to report bias (or prevent it)

Reward team members for finding ethics bugs

Dr. Ayanna Howard
- on the Artificial Intelligence Podcast with Lex Fridman

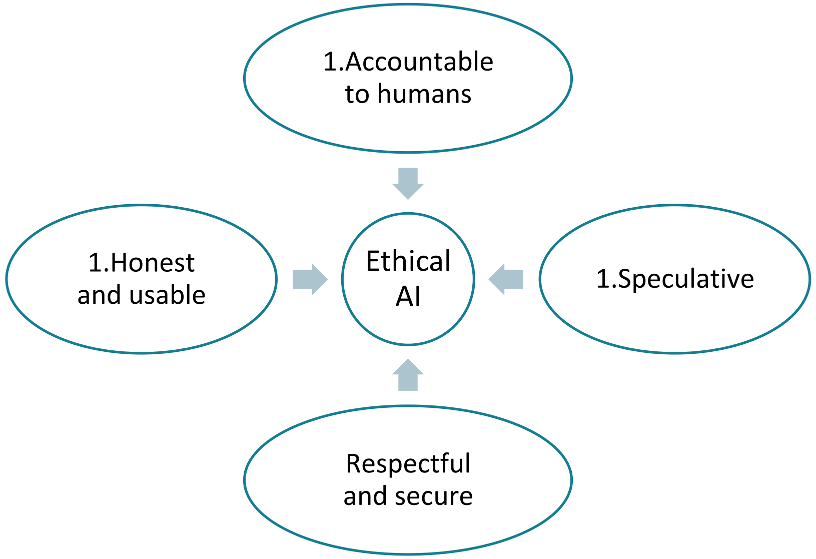
We aren’t perfect, AI won’t be perfect

Empower diverse teams, inclusive environments

Adopt technical ethics

Encourage deep conversations (Checklist)

Activate curiosity; be speculative; imaginative



Evangelize for human values

Ethical. Transparent. Fair.

Carol J. Smith

Twitter: @carologic

LinkedIn: https://www.linkedin.com/in/**caroljsmith**/

CMU’s Software Engineering Institute,
Emerging Technology Center

Twitter: @sei_etc